A study of analogical grids extracted using feature vectors on varying vocabulary sizes in Indonesian

Rashel Fam Yves Lepage
Graduate School of Information, Production, and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
{fam.rashel@fuji., yves.lepage@}waseda.jp

***Abstract*—Indonesian as an agglutinating language is known for its derivative morphological richness. Word forms are constructed by combining stem and affixes. In this paper, we study the influence of surface form and morphological information in analogical grids extracted from a set of word forms with varying sizes. Each word form is represented as a feature vector. In the experiment setting, we consider three features: characters, affixes, and morphosyntactic definition. The sizes and saturation are then observed to characterize the extracted grids.**

***Keywords*-analogy; analogical grid; word form; morphology**

I. Introduction

Analogical grids are tables containing word forms in its cells. They give a compact view on the organization of a lexicon up to a certain extent. Such grids are the result of an empirical procedure. Analogical grids are built from a set of word forms contained in a corpus. Thus, there are word forms which are not contained in the given corpus. These are called unseen word forms. This is the reason why we see empty cells inside analogical grids. They may be seen as a preliminary step towards the production of paradigm tables. In contrast with analogical grids paradigm tables are supposed to be completed. Paradigm tables are tables that show conjugation and declension of word forms in a language. They are usually found in dictionaries.

The left of Figure 1 is an example of an analogical grid in Indonesian. They can be used to study the productivity of a language [1], [2], [3]. [4] gave such analysis across 12 languages using analogical grids built from the Bible corpus [5]. As another example of use, [6] showed how to produce new word form in French by analogy from the neighbouring word forms in the same series. [7], [8] showed how to use analogical grid and analogy to perform word inflection task promoted in SIGMORPHON Shared Task [9], [10], [11].

Indonesian is an agglutinating language known for its morphological richness, mostly on derivational morphology. Word forms are constructed from stem and affixes (prefixes, suffixes and circumfixes). It also has reduplications, for example, *orang* 'person' $\rightarrow$ *orang-orang* 'people' in marked plural. In this paper, we study the analogical grids extracted from set of word forms contained in Indonesian morphological dictionary. We consider the use of three types of feature vectors: characters, affixes, and morphosyntactic definition to represent word forms. We analyse the influence of using different feature vectors on analogical grids size and saturation. Experiments also carried on varying sizes of vocabulary to observe the influence brought to the extracted grids.

This paper is organized as follows. Section II and III explains the notions of analogical grid and feature vectors for representing word forms. Section V and IV explains the experiment protocol and data used for the experiment. Section VI shows the results and analysis of it. Last, Section VII concludes the paper.

II. Analogical grid

Analogical grid is an M x N matrix, explained by the left side of Formula (1), where any four

makan	makanan	dimakan	memakan
minum	minuman	diminum	meminum
lihat		dilihat	
pukul	pukulan		
lirik	lirikan	dilirik	

makan : *makanan* :: *pukul* : *pukulan*

minum : *lirik* :: *diminum* : *dilirik*

minuman : *lirikan* :: *diminum* : *dilirik*

Figure 1. An analogical grid in Indonesian (*left*) and 3 out of many analogies that can be extracted from it (*right*).

objects contained in the cell on the same two columns and two rows is a proportional analogy between sequences of characters [12], [13] (see the right side of the Formula (1)).

$$\begin{array}{c} P_1^1 : P_1^2 : \cdots : P_1^m \\ P_2^1 : P_2^2 : \cdots : P_2^m \\ \vdots \quad \vdots \qquad\quad \vdots \\ P_n^1 : P_n^2 : \cdots : P_n^m \end{array} \overset{\Delta}{\Longleftrightarrow} \begin{array}{c} \forall (i,k) \in \{1,\ldots,n\}^2, \\ \forall (j,l) \in \{1,\ldots,m\}^2, \\ P_i^j : P_i^l :: P_k^j : P_k^l \end{array} \tag{1}$$

A proportional analogy (see Formula (2)) is defined as a relationship between four objects where two properties are met:

- equality of ratios (defined hereafter) between the first and the second terms on one hand, and the third and the fourth terms on the other hand, and
- exchange of the means (the second and the third terms can always be exchanged).

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \begin{cases} A : B &= C : D \\ A : C &= B : D \end{cases} \tag{2}$$

The right of Figure 1 shows several analogies that can be extracted from the analogical grid in Figure 1.

We define the ratio between two words in Formula (3) as a vector of features made up of the difference between feature vectors which represent the two word forms, plus, the distance between the two words.

$$A : B \overset{\Delta}{=} \begin{pmatrix} A_1 - B_1 \\ A_2 - B_2 \\ A_3 - B_3 \\ \vdots \\ A_n - B_n \\ d(A,B) \end{pmatrix} \tag{3}$$

In Formula (3), the notation S_c stands for the value of dimension c of feature vector representing string S. Section III explains this feature vector in detail. The last dimension, written as $d(A,B)$, is the edit distance[1] between the two strings. This indirectly gives the number of common characters appearing in the same order in A and B.

[4] proposed an algorithm to produce analogical grids from a set of words. All word pairs with the same ratio is grouped together as analogical clusters. Analogical grids are then constructed by adding these clusters as rows or columns into the grid.

A. Size and saturation of analogical grid

The size of an analogical grid is defined by the total number of cells inside the grid. To calculate the size of an analogical grid we can simply multiply the number of rows by its number of columns. Intuitively, analogical grid of word forms with bigger size means higher number of variation in a series of conjugation between word form.

Saturation is a way to measure how dense is a grid. The higher the saturation of analogical grid can be interpreted as how regular the conjugation exists in the language. Formula (4) shows how we calculate the saturation of an analogical grid.

$$\text{Saturation} = \frac{\text{Number of non-empty cells}}{\text{Total number of cells}} \times 100\% \tag{4}$$

Thus, the analogical grid in Figure 1 has size of $4 \times 5 = 20$ and saturation of $\frac{15}{20} \times 100\% = 75\%$.

III. Vector Representation for Word Form

Each word form in the list is converted into a feature vector to construct analogical grids.

[1] The only two edit operations used are insertion and deletion.

Feature vectors can be built from properties that define the word forms. This time, we consider three types of feature to represent word forms.

A. Characters as features

We use characters that occur in the word form as the feature of our vectors. The dimension of our feature vector is as long as the size of the alphabet in the language.

$$A = \begin{pmatrix} |A|_a \\ |A|_b \\ |A|_c \\ \vdots \\ |A|_z \end{pmatrix} \qquad \text{makanan} = \begin{pmatrix} 3 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5}$$

This is the basic feature vector and will be our baseline. Here, the notation $|A|_c$ stands for the number of occurrences of character c in string A. Thus, the word form *makanan* 'food' has a value of 3 for feature 'a' in the vector.

B. Affixes as features

Instead of only using the characters, we can also use affixes that construct the word form as features. This information can be obtained from stemmer or morphological analyser. In our experiments, we use manually checked affixes defined in MALINDO Morph[2] data set, instead of using a morphological analyser. It contained four kinds of affixes:

- prefix: *meN-*, *N-*, *di-*, etc.
- suffix: *-an*, *-kan*, *-i*, etc.
- circumfix: *ber- -an*, *ke- -an*, *peN- -an*, etc.
- reduplication: *full*, *partial*, and *rythmic*

For more examples of the affixes, please refer to [14]. In this setting, the labels are converted into Boolean values (*True*: 1 and *False*: 0).

$$A = \begin{pmatrix} -an \\ di- \\ meN- \\ \vdots \\ pe- \end{pmatrix} \qquad \text{makanan} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{6}$$

[2]github.com/matbahasa/MALINDO_Morph

C. Morphosyntactic definitions as features

Previous settings focus only on the level of surface form. It is also possible to use morphosyntactic definitions as features. Such feature vector can be built, for instance, from the Unimorph Project [15] data which have been built from parsing Wiktionary data into a language-independent feature schema [16], [17]. Morphosyntactic definitions may consist of tense, case, gender, number, part-of-speech tag, etc. However, for some low-resource languages, the number of word forms and morphosyntactic definition of word forms contained in the data is very small.

$$A = \begin{pmatrix} NND \\ NSD \\ ADJ \\ \vdots \\ VSA \end{pmatrix} \qquad \text{makanan} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7}$$

In our experiments, we only consider the part-of-speech tag retrieved from morphological analyser for Indonesian, MorphInd[3] [18], as morphosyntactic features. Similar to the previous setting, the labels are Boolean. For example, the word form *makanan* 'food' is a noun so that its value for feature 'NND' is 1 (True).

IV. Data used

We use MALINDO Morph as our dataset. It is a morphological dictionary for Malay/Indonesian. Figure 2 shows an example of entries in MALINDO Morph data. Its entry forms are collected from authoritative dictionaries in Malaysia (*Kamus Dewan*) and Indonesia (*Kamus Besar Bahasa Indonesia*). It also contained some expanded entries from Leipzig Corpora Collection. Most of the entries of the data are checked by hand. Table I gives statistics on the original MALINDO Morph data.

Although both Malay and Indonesian are mutually intelligible, there are some variations exist between the two languages. In this paper, we focus only on Indonesian. Unfortunately, no information

[3]github.com/neocl/morphind

Root	Surface form	Prefix	Suffix	Circumfix	Reduplication
perlu	perlu	0	0	0	0
perlu	seperlunya	0	0	se- -nya	0
perlu	memerlukan	meN-	-kan	0	0

Figure 2. An example of entries in MALINDO Morph data taken from [14]. '0' means there is no corresponding affix contained in the respective surface form.

Table I
STATISTICS ON THE MALINDO MORPH DICTIONARY

Number of tokens	233,372
Average length of tokens	8.14±3.01
# prefix types	12
# suffix types	10
# circumfix types	7
# reduplication types	3

in the dataset explains whether a word form is Malay, Indonesian, or both. Therefore, we rely on an Indonesian morphological analyzer, MorphInd, to *recognize* the word form.

The surface forms and root is lowercased before filtered against MorphInd. We consider only word forms that can be parsed by MorphInd (MorphInd did not give *X* 'unknown' or *F* 'foreign word' as its output). After that, we randomly choose word forms to construct the analogical grids with varying vocabulary sizes: 100, 500, 1000. Tokens in the dataset with smaller sizes are always included in the dataset with bigger sizes.

V. EXPERIMENT PROTOCOL

Experiments are carried on Indonesian word forms contained in the MALINDO Morph data. Analogical grids are constructed using different types of feature vector on varying vocabulary sizes. We then measure the sizes and saturations of the extracted grids.

VI. RESULTS AND ANALYSIS

Table II shows the result of analogical grids extracted using different feature vectors on varying vocabulary sizes (number of word forms), where:

- CHAR: characters feature,
- AFF: affixes feature, and
- MORPH: morphosyntactic definition.

The notation '+' means that the feature vector is added up. For example, CHAR+AFF use both characters and affixes as features.

A. Number of analogical grids

Results show that CHAR struggles to build any analogical grid on the smaller sizes of vocabulary settings. This may be caused by less number of word forms where analogies hardly emerge among the word forms. CHAR+MORPH and CHAR+AFF+MORPH give the highest number of grids. The higher the number of word forms used to construct the analogical grids the higher the number of analogical grids produced. We can see a huge increase in the number of analogical grids under smaller vocabulary size. For example, an increase up to 10 times can be observed on vocabulary size of 100 to 500 word forms before dropping to only around 2 times on vocabulary size of 500 to 1,000 word forms. The number of analogical grids can be interpreted as a rough estimation on how rich is the variation of conjugations in a language.

B. Average size of analogical grids

The average sizes of analogical grids show a stable increase with bigger vocabulary size. CHAR+MORPH feature vector always gives smaller numbers than the other combinations. This may due to the heavier constraint on morphological feature rather than using AFF feature. It forces the grid to limit similar part-of-speech tag changes.

C. Average saturation of analogical grids

In contrast with the previous observations, we can see that the saturation of analogical grids seems to be very similar although being produced from different feature vectors. Results also show that bigger vocabulary size leads to decreasing

Table II
ANALOGICAL GRIDS EXTRACTED FROM DIFFERENT FEATURE VECTORS ON VARYING VOCABULARY SIZES

Feature	Vocabulary size	Number of analogical grids	Average size	Average saturation (%)
CHAR	100	0	-	-
	500	0	-	-
	1,000	9	10	94.18
CHAR+AFF	100	9	8	100.00
	500	39	395	92.88
	1,000	86	2,713	88.24
CHAR+MORPH	100	7	8	98.12
	500	69	221	92.17
	1,000	143	2,080	89.87
CHAR+AFF + MORPH	100	7	8	98.14
	500	69	238	93.54
	1,000	139	2,247	89.78

saturation. This meets intuition because the growth of grid size is not equal with the growth of the number of word forms inside the analogical grid.

VII. CONCLUSION

We constructed analogical grids from different feature vectors: characters, affixes, and morphosyntactic definitions of word form. Results show that the use of richer feature vector, both surface form and morphological information, put more constraint on the analogical grids.

For other languages where there is no such morphological dictionary, we may consider the use of unsupervised approach to learn the affixes, like [19]. Filling the empty cells inside the grids and checking the validity of those newly generated word form should be performed, like [20].

REFERENCES

[1] R. Singh and A. Ford, "In praise of Sakatayana: some remarks on whole word morphology," in *The Yearbook of South Asian Languages and Linguistics-200*, R. Singh, Ed. Thousand Oaks: Sage, 2000.

[2] S. Neuvel and S. A. Fulop, "Unsupervised learning of morphology without morphemes," in *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*. Association for Computational Linguistics, July 2002, pp. 31–40. [Online]. Available: http://www.aclweb.org/anthology/W02-0604

[3] N. Hathout, "Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy," in *Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*. Manchester, UK: Coling 2008 Organizing Committee, August 2008, pp. 1–8. [Online]. Available: http://www.aclweb.org/anthology/W08-2001

[4] R. Fam and Y. Lepage, "Morphological predictability of unseen words using computational analogy," in *Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-CA-16)*, Atlanta, Georgia, 2016, pp. 51–60.

[5] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: the bible in 100 languages," *Language Resources and Evaluation*, vol. 49, no. 2, pp. 375–395, Jun 2015. [Online]. Available: https://doi.org/10.1007/s10579-014-9287-y

[6] N. Hathout, "Acquisition of morphological families and derivational series from a machine readable dictionary," *CoRR*, vol. abs/0905.1609, 2009. [Online]. Available: http://arxiv.org/abs/0905.1609

[7] R. Fam and Y. Lepage, "A holistic approach at a morphological inflection task," in *Proceedings of the 8th Language and Technology Conference (LTC-17)*. Poznań, Poland: Fundacja uniwersytetu im. Adama Mickiewicza, November 2017, pp. 88–92.

[8] R. Fam and Y. Lepage, "IPS-WASEDA system at CoNLL–SIGMORPHON 2018 Shared Task on morphological inflection," in *Proceedings of the*

CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. Brussels: Association for Computational Linguistics, October 2018, pp. 33–42. [Online]. Available: https://www.aclweb.org/anthology/K18-3003

[9] R. Cotterell, C. Kirov, J. Sylak-Glassman, D. Yarowsky, J. Eisner, and M. Hulden, "The SIGMORPHON 2016 Shared Task—Morphological Reinflection," in *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, 2016, pp. 10–22. [Online]. Available: http://aclweb.org/anthology/W16-2002

[10] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, P. Xia, M. Faruqui, S. Kübler, D. Yarowsky, J. Eisner, and M. Hulden, "CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages," in *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Vancouver: Association for Computational Linguistics, August 2017, pp. 1–30. [Online]. Available: http://www.aclweb.org/anthology/K17-2001

[11] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden, "The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection," in *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*. Association for Computational Linguistics, 2018, pp. 1–27. [Online]. Available: http://aclweb.org/anthology/K18-3001

[12] P. Langlais and F. Yvon, "Scaling up analogical learning," in *Coling 2008: Companion volume: Posters*. Manchester, UK: Coling 2008 Organizing Committee, August 2008, pp. 51–54. [Online]. Available: http://www.aclweb.org/anthology/C08-2013

[13] N. Stroppa and F. Yvon, "An analogical learner for morphological analysis," in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 120–127. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0616

[14] D. M. Hiroki Nomoto, Hannah Choi and F. Bond, "MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, K. Shirai, Ed. Paris, France: European Language Resources Association (ELRA), May 2018.

[15] C. Kirov, J. Sylak-Glassman, R. Que, and D. Yarowsky, "Very-large scale parsing and normalization of wiktionary morphological paradigms," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), May 2016.

[16] J. Sylak-Glassman, C. Kirov, D. Yarowsky, and R. Que, "A language-independent feature schema for inflectional morphology," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 674–680. [Online]. Available: http://www.aclweb.org/anthology/P15-2111

[17] J. Sylak-Glassman, C. Kirov, M. Post, R. Que, and D. Yarowsky, *A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging*. Cham: Springer International Publishing, 2015, pp. 72–93. [Online]. Available: https://doi.org/10.1007/978-3-319-23980-4_5

[18] S. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (MorphInd): Towards an Indonesian corpus," *Systems and Frameworks for Computational Morphology*, pp. 119–129, 2011.

[19] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Trans. Speech Lang. Process.*, vol. 4, no. 1, pp. 3:1–3:34, Feb. 2007. [Online]. Available: http://doi.acm.org/10.1145/1187415.1187418

[20] R. Fam, A. Purwarianti, and Y. Lepage, "Plausibility of word forms generated from analogical grids in Indonesian," in *Proceedings of the 16th International Conference on Computer Applications (ICCA-18)*, Yangon, Myanmar, February 2018, pp. 179–184.